ABSTRACT OF THE DISCLOSURE

A method and apparatus are provided for determining when electronic documents stored in a large collection of documents are similar to one another. A plurality of similarity information is derived from the documents. The similarity information may be based on hyperlinks in the documents, text similarity, similarity of multimedia components in the documents, user click-through information, similarity in the titles of the documents or their location identifiers, etc. Another source of similarity information is patterns of user viewing, as may be monitored by a Web caching system. A pair of documents may be inferred to be similar if a particular user has shown high interest in both of them within a particular session or time period; alternatively, a pair of documents may be inferred to be similar if the pattern interest in the by all users for all sessions is similar. User interest is considered to be a function of the time that a user has viewed the document. The similarity information is fed to a combination function that synthesizes the various measures of similarity information into combined similarity information. Using the combined similarity information, an objective function is iteratively maximized in order to yield a generalized similarity value that expresses the similarity of particular pairs of documents. In an embodiment, the generalized similarity value is used to determine the proper category, among a taxonomy of categories in an index, cache or search system, into which certain documents belong.